Enabling Language Models to Implicitly Learn Self-Improvement

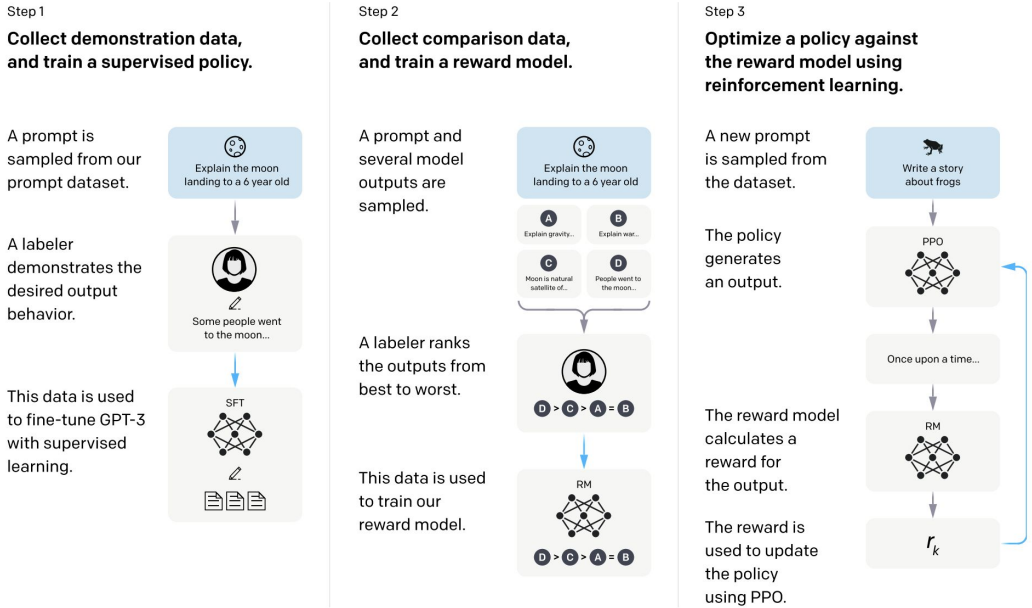
Ziqi Wang, Le Hou, Tianjian Lu, Yuexin Wu, Yunxuan Li, Hongkun Yu, Heng Ji

Motivation: Enabling models to self-improve

Preference data contains self-improvement signals!

Preference Data:

1. Contains a reference response and an improved response
2. Implicitly contains the self-improvement information



[1] Ouyang, Long, et al. "Training language models to follow instructions with human feedback." *Advances in Neural Information Processing Systems* 35 (2022): 27730-27744.

Motivation: Enabling models to self-improve

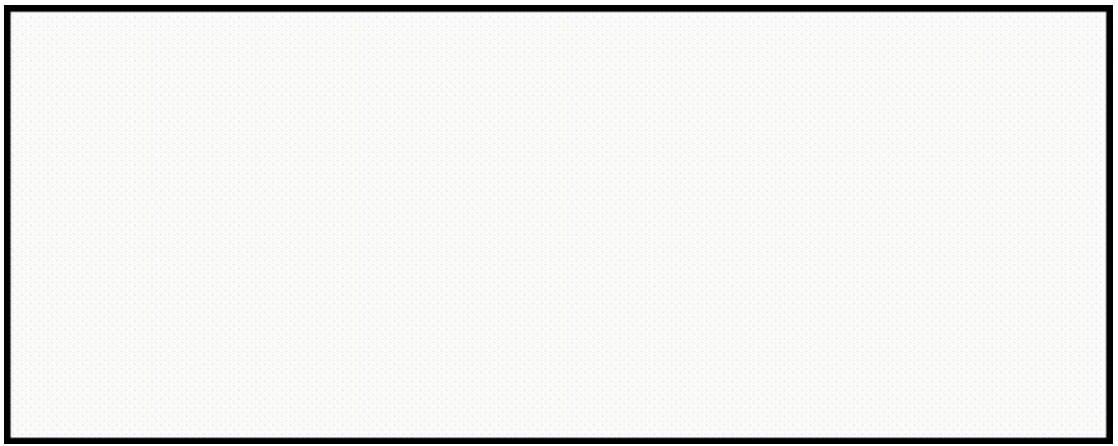
Preference data can not only be used to train LLMs that generate **good** responses, but can also train models that generate **better** responses based on reference responses

$$y_{\mathrm{P}} \sim \mathbf{M}_{\mathrm{P}}(\cdot|x) \longrightarrow y_{\mathrm{PIT}} \sim \mathbf{M}_{\mathrm{PIT}}(\cdot|x, y_{\mathrm{ref}})$$

Why not prompting for self-improvement?

Prompting: “Can you reflect your response and improve it?”

Improve:
- ➔ More helpful?
- ➔ More formal?
- ➔ More specific?
- ➔ More objective?



More helpful?
- ➔ More polite
- ➔ More details
- ➔ Avoid irrelevant info
- ➔ Avoid answering dangerous questions
- ➔ ……

[1] Madaan, Aman, et al. "Self-refine: Iterative refinement with self-feedback." arXiv preprint arXiv:2303.17651 (2023).

Why not prompting for self-improvement?

Human: What is the origin of the name "Black Friday"?

PaLM 2: ... The first stores to hold such sales in the 1950s called them "Black Friday" because of the increased profits they experienced.

Human: Can you reflect your response and give a more helpful response to the question?

PaLM 2: .....Shopping on Black Friday can be stressful and uncomfortable for many people due to the crowds and crowds, and can also result in overspending and impulse purchasing. For many consumers, the stress of Black Friday shopping outweighs the potential financial benefits, and so shopping on Black Friday is avoided by many consumers.

Why not prompting for self-improvement?

It is generally **hard** to describe the self-improvement requirements precisely **in language**.

Learning from preference data is an **easier** way since models can extract the self-improvement requirements implicitly from data.

Methods

$$y_{\mathrm{P}} \sim \mathbf{M}_{\mathrm{P}}(\cdot|x) \longrightarrow y_{\mathrm{PIT}} \sim \mathbf{M}_{\mathrm{PIT}}(\cdot|x, y_{\mathrm{ref}})$$

Reformulate RLHF pipeline:

- ➔ **Supervised Fine-Tuning**
- ➔ **Reward Model Training**
- ➔ **Reinforcement Learning**
- ➔ **Inference**

Method

➔ **Supervised Fine-tuning:** Learn to generate **<better>** human written responses

Input x

Should I buy a convertible car?

Response y

It depends .....

$$-\sum_{(x,y_l,y_w)\in\mathcal{D}_{\text{SFT}}} \log \mathrm{M_P}(y_w|x)$$

Input x

Should I buy a convertible car?

Worse Response y_l

Yes.

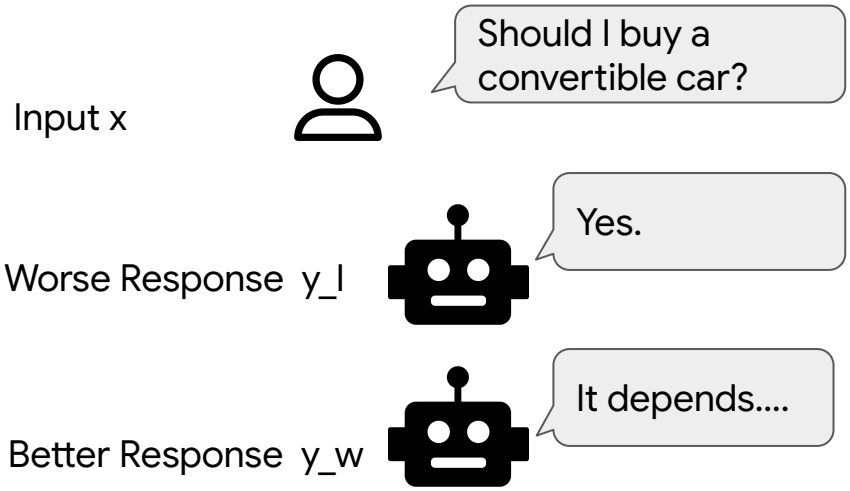
Better Response y_w

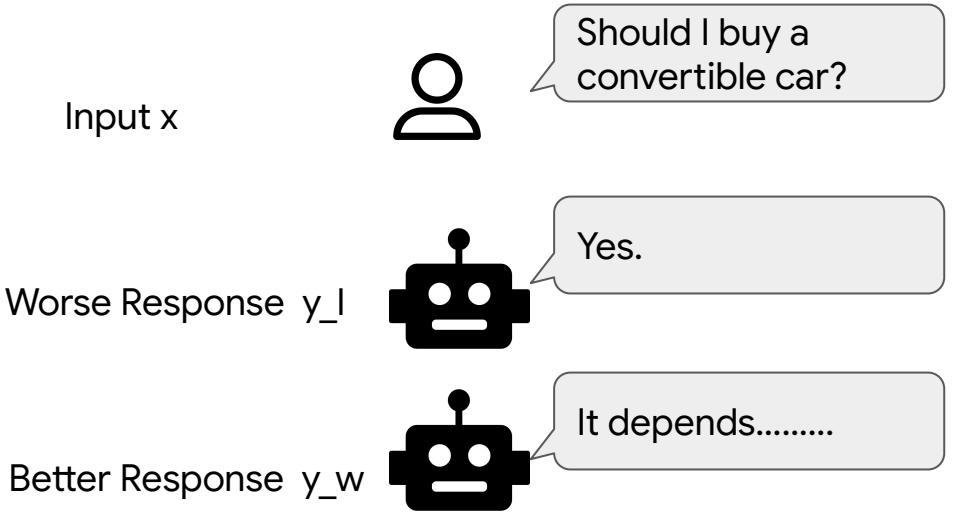
It depends.....

$$-\sum_{(x,y_l,y_w)\in\mathcal{D}_{\text{SFT}}} \log \bar{\mathrm{M}}_{\text{PIT}}(y_w|x,y_l)$$

Method

➔ **Reward Model Training:** Learn to distinguish between **<better>** and **<worse>** responses



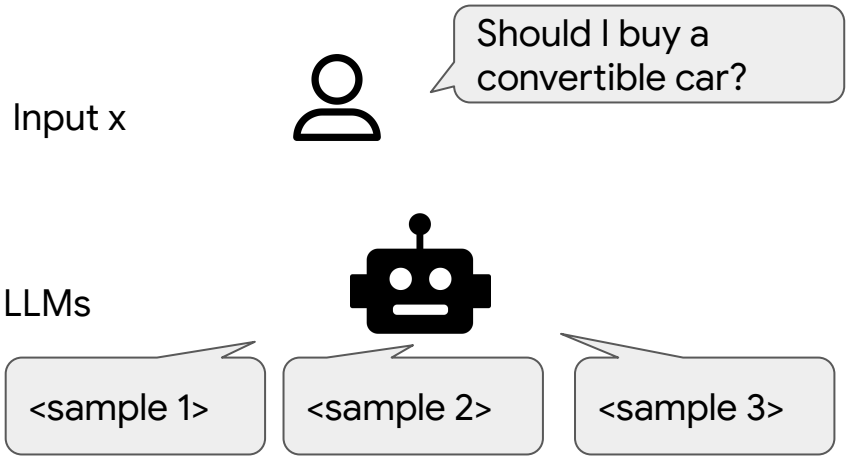


$$-\sum_{\mathcal{D}_{\text{RM}}} \log \sigma(r_w - r_l)$$

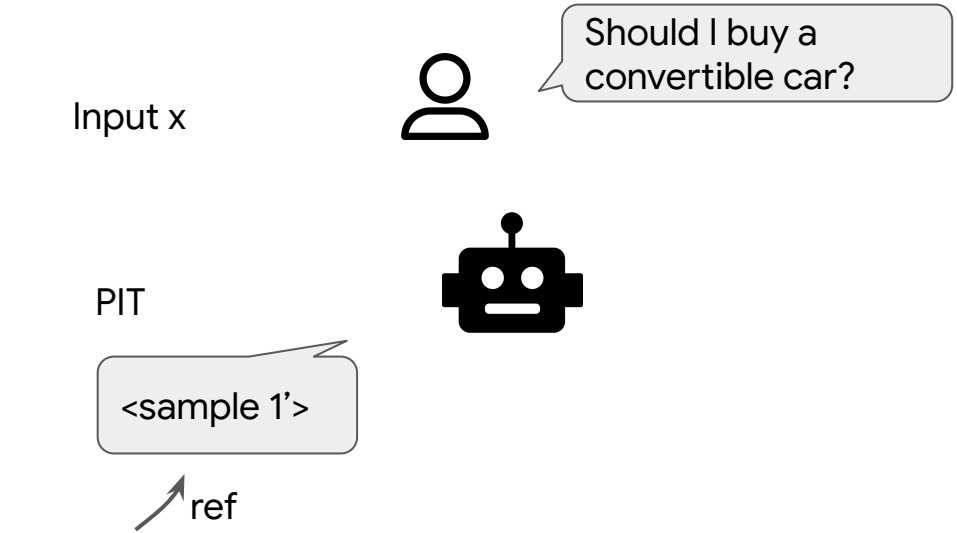
$$r_{\text{gap}}(x, y_w, y_l) \geq r_{\text{gap}}(x, y_w, y_w)$$
$$\approx r_{\text{gap}}(x, y_l, y_l) \geq r_{\text{gap}}(x, y_l, y_w)$$

Method

➔ **Reinforcement Learning:** Learn to generate **<better>** responses



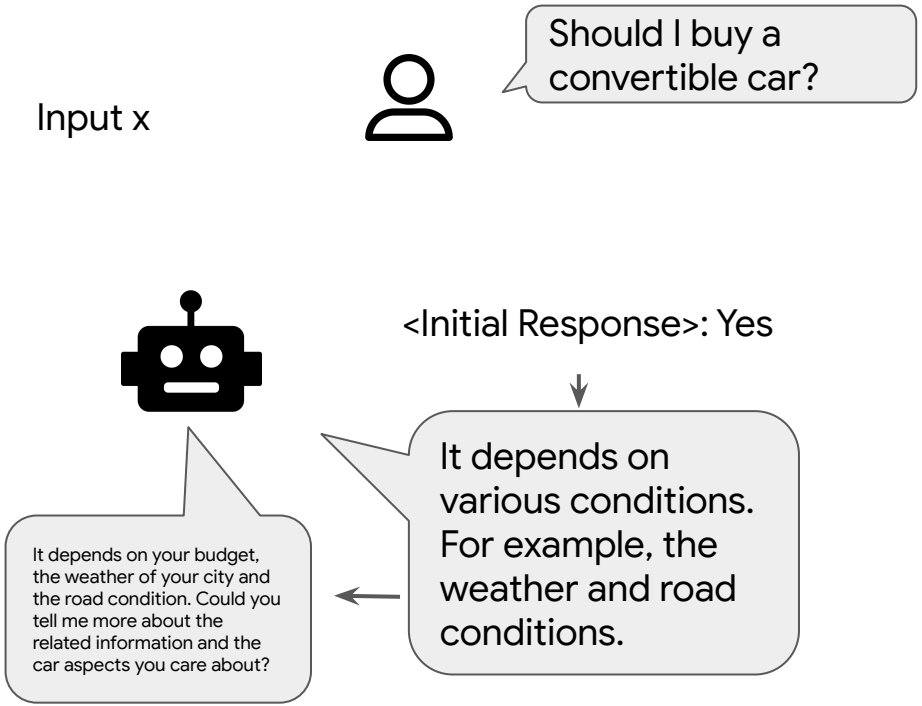
$$\sum_{\mathcal{D}_{\text{RL}}} \left[r(x, y) - \beta \mathbf{KL}(\mathbf{M}_{\mathbf{P}}^{\text{RL}}(y|x) - \mathbf{M}_{\mathbf{P}}^{\text{SFT}}(y|x) \right]$$



$$\sum_{\mathcal{D}_{\text{RL}}} \sum_{y_{\text{ref}} \in \{y_l, y_w\}} \left[r_{\text{gap}}(x, y, y_{\text{ref}}) - \beta \mathbf{KL}(\mathbf{M}_{\text{PIT}}^{\text{RL}}(y|x, y_{\text{ref}}) - \mathbf{M}_{\text{PIT}}^{\text{SFT}}(y|x, y_{\text{ref}}) \right]$$

Method

➔ **Inference:** Generate a **<better>** response



Experiment: PIT outperforms Self-Refine

1. PIT improves the LLM response.
2. Self-Refine improves the LLM response, too.
3. PIT performs better than Self-Refine.

Dataset	Comparison	Win rate / Lose rate / Δ (%) GPT-4	DeBERTa	Human Evaluation
Anthropic/HH-RLHF	Original vs. y_w	71.85/17.19/54.69	68.20/18.00/50.20	-
	PIT vs. Original	55.47/27.34/28.13	46.30/32.30/14.00	-
	Self-Refine vs. Original	60.94/17.19/43.75	40.30/31.40/8.90	-
	PIT vs. Self-Refine	38.28/42.19/−3.91	41.3/37.60/3.70	47.06/23.53/23.53
OpenAI/Summary	Original vs. y_w	74.22/8.59/65.63	84.90/10.70/74.20	-
	PIT vs. Original	44.53/24.22/20.31	41.9/34.7/7.2	-
Synthetic Data	Original vs. y_w	28.91/51.56/−22.66	-	-
	PIT vs. Original	48.44/14.84/33.59	-	-
	Self-Refine vs. Original	34.38/17.97/16.41	-	-
	PIT vs. Self-Refine	45.31/35.16/10.16	-	-

Conclusion

$$y_{\mathrm{P}} \sim \mathbf{M}_{\mathrm{P}}(\cdot|x) \longrightarrow y_{\mathrm{PIT}} \sim \mathbf{M}_{\mathrm{PIT}}(\cdot|x, y_{\mathrm{ref}})$$

Thank you!

Paper: https://arxiv.org/pdf/2310.00898.pdf